Predictive Analysis to Improve Crop Yield using a Neural Network Model

Shruti Kulkarni
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology,
Bangalore, India

Shah Nawaz Mandal
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology,
Bangalore, India

G Srivatsa Sharma
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology,
Bangalore, India

Monica R Mundada
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology,
Bangalore, India

Meeradevi
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology,
Bangalore, India

***Abstract*— Agriculture has been the sector of paramount importance as it feeds the country population along with contributing to the GDP. Crop yield varies with a combination of factors including soil properties, climate, elevation and irrigation technique. Technological developments have fallen short in estimating the yield based on this joint dependence of the said factors. Hence, in this project a data-driven model that learns by historic soil as well as rainfall data to analyse and predict crop yield over seasons in several districts, has been developed. For this study, a particular crop, Rice is considered. The designed hybrid neural network model identifies optimal combinations of soil parameters and blends it with the rainfall pattern in a selected region to evolve the expectable crop yield. The backbone for the predictive analysis model with respect to the rainfall is based on the Time-Series approach in Supervised Learning. The technology used for the final prediction of the crop yield is again a branch of Machine Learning, known as Recurrent Neural Networks. With two inter-communicating data-driven models working at the backend, the final predictions obtained were successful in depicting the interdependence between soil parameters for yield and weather attributes.**

***Keywords*— Activation function, Crop yield, Hybrid model, Joint prediction, Machine Learning, Neural Networks, Time series**

I. Introduction

Agriculture is the backbone of Indian economy. The yield obtained primarily depends on weather conditions as rainfall patterns largely influence cultivation methodologies. With this context, farmers and agriculturalists require spontaneous advice proposition in predicting future reaping instances to maximize crop yield.

Due to insufficient involvement of technology, the throughput of agriculture is yet to reach its full glory. Every farmer is interested in knowing the yield he/she could expect at the harvest period and hence, yield prediction is an important aspect for them. Over the years, farmers have an idea about the pattern in yield as per innate human intuition. However, rainfall as a major driver for crop raising can extensively rattle intuitive yield prediction by controlling some of the soil and environmental parameters related to the crop growth. Moreover, the right kind of soil to be employed for a crop is only known to the farmer only by on-paper advice and makes it difficult for him/her to trial and test on crop investment.

The proposed architecture provides a computational dimension to enhance knowledge about the yield before the crop sowing period. It is made possible through a data driven hybrid model. Since the model performs a joint prediction of both rainfall and soil features on the yield, it is termed as a hybrid model. This model reads from data of past period and learns through computational processes like ARIMA, Exponential Sequencing and Recurrent Neural Networks. The model accepts soil features, fertilizer concentration and sowing period from the end user as test data for the trained model. The output is displayed as the expected yield in tonnes per hectare, for the given parameters. The soil features include the concentration of Nitrogen, Phosphorous and Potassium in the soil and fertilizer, and the mean pH value of the soil.

The rainfall analysis model forecasts rainfall quantity in future time by learning and analyzing huge sets of past data, using ARIMA and Exponential Sequencing. On the other hand, a Recurrent Neural Network(RNN) is constructed for analyzing and learning crop yield patterns on the soil types. The RNN also develops predictions on the test data using the results of rainfall predictions from the previous model flow. Thus, the proposed approach boils up a joint rainfall-soil-yield prediction for futuristic knowledge about crop sowing period and yield expectations. The approach not only promises precise levels of expected yield but also helps determine seasonality trends in

978-1-5386-5314-2/18/$31.00 ©2018 IEEE

gauging the yield of the same crop over different years in the same season. This approach is highly beneficial to farmers, agriculturalists, local self-governments, and Tahsildars to observe and allocate capital for agriculture and crop raising. The proposed approach also tries to solve the misery of high suicide rate of farmers in the country.

II. Literature Survey

The research by Rasul G et al. [1] is about how the unexpected rise in temperature in Pakistan has led to the crop yield getting disturbed. Further, the paper reveals how this rise has induced a negative impact on the crop yield, which in turn is leading to country wide agricultural issues.

Anupama Mahato [2] in her study, analysed the long term climatic change that could affect agriculture in several ways. Some of the ways discussed are about the quantity and quality of crops in terms of productivity, growth rates, photosynthesis and transpiration rates, moisture availability, etc. Also, in her paper, she has presented the impact of climatic change on the global food production that can threaten the food security levels.

The research by Japneet Kaur [3] is based on state level data of four major seasonal Indian crops such as Rice, Wheat, Cotton, Sugarcane which comprise of Food and Cash crops for the time span of 2004 to 2013. Seven agriculturally intensive states with varied climatic conditions have been taken into consideration for this study that makes an attempt to analyse the impact of climate change on Indian Agriculture and food security.

The research study by Pratap S. Birthal et al. [4] examined the impact of climate change in a country like India that have limited arable land but heavy dependence on agriculture. It shows the inadequate technological and financial capabilities for mitigation and adaptation to climate change. The paper reveals that in the recent years, the climate change has been accompanied by increased incidence of natural calamities such as droughts, cyclones and floods. The study concludes with the proof that such events can cause a drastic decline in the agricultural output, exacerbating the problems of food insecurity and rural poverty.

Yunis H et al. [5] correlates field weather and bacterial speck of tomato. This paper showed how the plant bacteria grows in a favorable range of temperature and hence, spoils the yield. The study said that the plant disease develops and spreads only at temperatures between 13 and 28 C and at relatively high humidity with free water on the leaves. Yield losses varied from 75% in plants infected at an early stage of growth to 5% in plants infected later during the season.

The research analysis made by J. P. Powell and S. Reinhard [6] showed that while the number of days with high temperatures in Dutch's wheat growing areas has evidently increased since the early 1900s, the number of low temperature days has dropped. The effects of weather conditions on the yield were diagnosed to be time specific. High temperature and precipitation together, were found to significantly decrease the amount of yield obtained.

III. Methodology

A. Dataset description

There are two data sets used for the hybrid model. The first, contains historic district-wise rainfall data for all the districts of Karnataka. The collection period spans to 30 years from 1987 to 2018. Rainfall is measured in millimetres and the labelled volume for a District is the mean of values recorded at all the weather stations in the District.

The other data set contains a detailed description about the soil properties recorded in all the Districts of Karnataka recorded over 31 years. Soil properties include the concentration of Nitrogen, Phosphorous and Potassium (NPK) in the soil (all in tonnes), the scales of pH of the soil, amongst others. Every row of values is labelled with a corresponding Yield value expressed in tonnes per hectare.

The hybrid model proposed in this paper curates results of the model trained on rainfall data with the RNN model trained on other soil properties.

B. Rainfall analysis using Time Series approach

Zhang [7] has evolved the use of time series analysis using hybrid ARIMA (Auto-Regressive Integrated Moving Average) methods. Time series analysis comprises of methods to analyse time series data in order to extract and develop characteristics from the data. Rainfall over seasons in a year tends to repeat over the following years. So, Rainfall data exhibits a natural temporal ordering.

This paper has utilised the data set to evolve a rainfall forecasting model using an open source forecasting tool, fbProphet [15] by Facebook. The approach of a time-series analysis includes taking into account the seasonality and trend that is observed in the rainfall pattern. A simple linear regression mostly fails to capture this trend observed and thereby, results in predictions with poor accuracy. This limitation is overcome with the help of an auto-regressive module fbProphet that ensures taking into account the variation of rainfall over the different seasons that a year witnesses [16]. This utility is used for fitting input datasets of multiple dimensions to a model, specifying the trend, variation and seasonality features (if known). fbProphet internally runs through techniques such as ARIMA [17] and Exponential Smoothening [18] on auto-fit data dimensions. The trends in annual rainfall (in mm) over a particular area can be visualised in Figure 1. The fbProphet modelling has been able to successfully train on the dataset fed to it, to fit test data points, ignoring the outliers. The model is also able to forecast future volume of rainfall from the present.

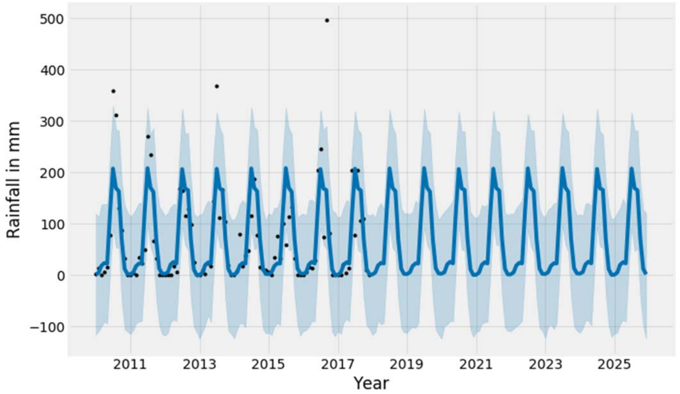

Fig. 1 An auto data-fit plot of rainfall volume until present, and the future forecasts for 7 years

The rainfall prediction modelled using fbProphet helps to learn input rainfall data through auto dimensioned dataframes. The so learned features in the dataframes help to develop data seasonality and regressive trends that proved to be unique for fbprophet. The modelling terminates with a dataframe formation with populated futuristic rainfall predictions. The prediction dataframes are employed for joint prediction analysis with respect to the influence of soil properties on crop yield.

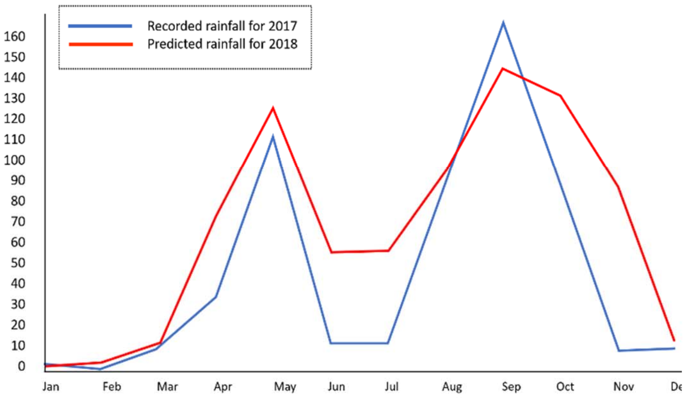

Fig. 2 A plot of recorded rainfall values in mm in 2017 over the year versus the predicted rainfall through the year 2018

C. *Recurrent Neural Networks*

Neural Networks in Machine Learning can be interpreted as a connection of nodes (or neurons) to form directed graphs through which data flows across layers. Like biological neural networks in the brain, Neural Networks have large number of perceptrons or neurons that are organised into several layers. Each layer may have variable number of neurons. The input layer neurons (similar to i0 and i1 in Fig. 3) interface with the input data set. The output layer neurons (similar to o0 in Fig. 3; In the proposed algorithm, there is only 1 neuron present at the output layer) pull out the predicted value for the input. Layers lying between the input and output layers are hidden layers. Each neuron in this layer (similar to h0,h1,h2,c0,c1 and c2 in Fig. 3) has a typical characteristic in the layer which it belongs to. Neurons exhibit characteristics though Activation Functions. Activation functions typically used in regression problems (in our case) are Linear, reLu, tanH and Logistic activation functions. Data flows from the input layer to the output layer through the hidden layers. At each layer, a new set of characteristics are learned by the Neural Network Model.

A class of Neural Networks are Recurrent Neural Networks wherein connections between neurons form a directed graph along a sequence. This property of a Recurrent Neural Network(RNN) enables it to exhibit a dynamic temporal behaviour for a time sequence.

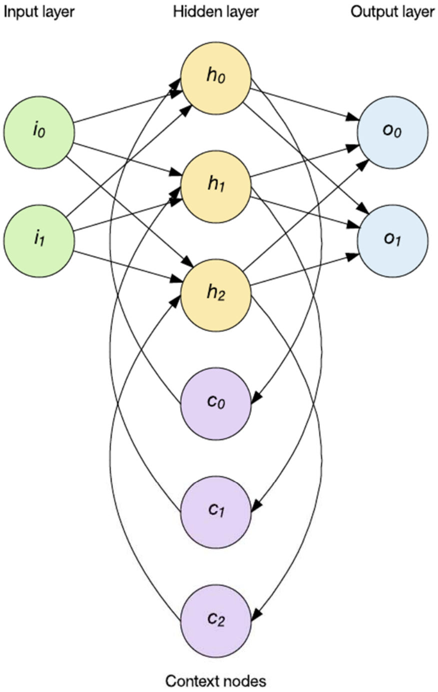

Fig. 3 Graphical form of a typical Recurrent Neural Network depicting nodes of different layers connected with directed edges

D. *The Recurrent Neural Network for soil feature modelling*

Neural Nets or machine learning cells have been evident in being able to revisit past patterns in the dataset. The features of the dataset under consideration include:

- Nitrogen, Phosphorous, Potassium concentration (NPK) in tonnes per hectare.
- pH grades of DMPH5, DMPH6, DMPH7, DMPH8 and DMPH9 (increasing order of basicity).
- Yield of Rice (Qrice) in Tonnes per Hectare.

- Rainfall (monthly as well as annual rainfall in mm) that resulted in the amount of yield, as recorded for any particular year in the dataset.

For every tuple of NPK and pH, Qrice is labeled as the yield function.

The architecture of the RNN model includes three layers. The input layer absorbs the above said features while the output layer pops out the labeled yield values. The input layer consists of 32 neurons that read NPK and pH values given by the user in the GUI, as the inputs. The reLu (rectified linear unit) activation function is used for the input layer neurons to act on the input provided. A hidden layer is developed with 16 neurons and is fully connected with the input layer. The hidden layer neurons are supplied with a 'Linear' activation function. The output layer neuron that is activated by a reLu function for 16 input pulses forms the final prediction. Before finishing up, this prediction made is displayed on the GUI designed.

The soil data set is split into training and test portions. 80% of the data is reserved for training the RNN model and the rest 20% is used for testing the performance of prediction of the model. Performance of the model is gauged using standard metrics which is discussed later in the paper.

The RNN model is tuned with a batch size of 4 and is run for 215 epochs. Thus there are 215 * 4 = 860 number of iterations over the training set for the considered hyperparameters (batch size and number of epochs). One epoch is one forward pass and one backward pass of all the training examples.

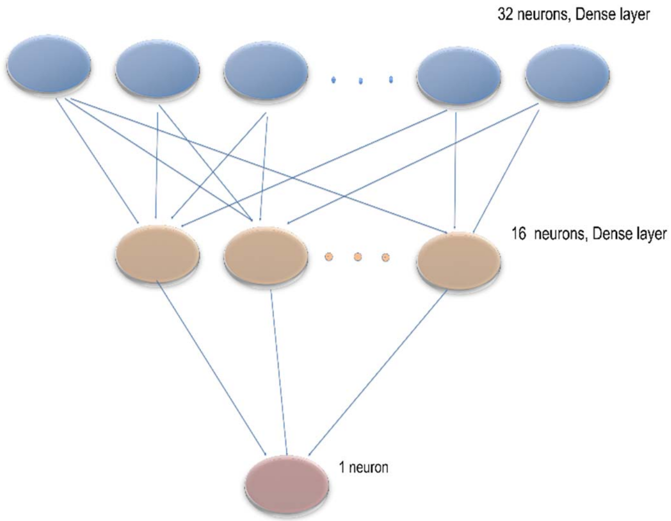

Fig. 4 Layered Architecture of the Neural Nerwork

IV. Algorithmic Details

The algorithm initially reads rainfall data using fbProphet data frames. Transformed data is compiled to extract patterns in monthly rainfall over several years sequentially. Using time-series forecasting, future values from the present can be predicted. The prediction model overlooks outliers and fits a recurring curve for past and futuristic data. During compilation of rainfall for a future interval of time, parallel reading of soil-yield data starts at the RNN. Data features are read into a Pandas (a Python programming library) dataframe into the computer memory. Training of the RNN models begins with the first 80% of the data tuples (rows) over the above described 215 epochs. At the completion of the training process, predicted values of yield (say rice) is compared with the recorded values from the dataset. The RNN model is validated by knowing the error in prediction vs true value. Over each epoch on the training samples, the loss (or cost) function gradually keeps reducing, indicating an increase in accuracy performance of the model.

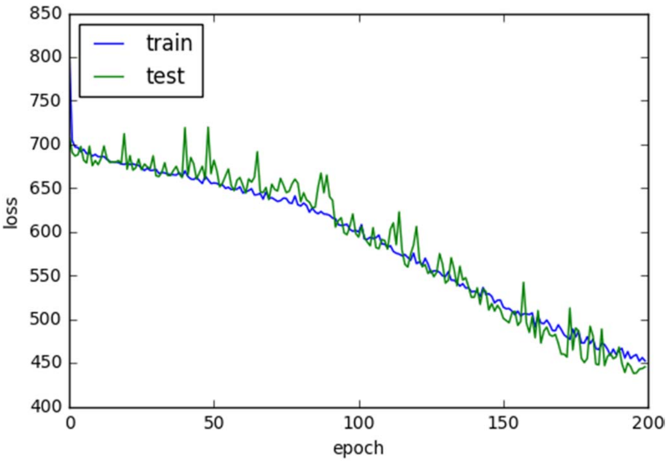

Fig 5 Reduction in Training and Testing loss over successive epochs

The completion of training and testing of the RNN model marks the reintroduction of the predictions obtained from fbProphet. Rainfall patterns in future time are synchronized with predicted values for soil properties. This methodology results in a hybrid model that jointly predicts the soil properties with the influence of rainfall on the prediction of the yield for Rice crop.

Figure 6 depicts the data flow involved in this algorithm discussed.

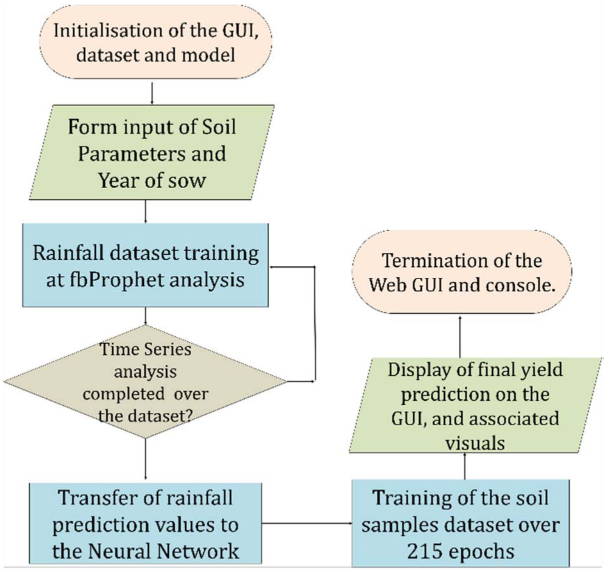
Initialisation of the GUI, dataset and model
Form input of Soil Parameters and Year of sow
Rainfall dataset training at fbProphet analysis
Time Series analysis completed over the dataset?
Transfer of rainfall prediction values to the Neural Network
Training of the soil samples dataset over 215 epochs
Display of final yield prediction on the GUI, and associated visuals
Termination of the Web GUI and console.

Fig 6. Data flow through the algorithm for the hybrid model

It was observed from the trends in data that crop yield in tonnes per hectare per annum is proportional to the annual rainfall quantity in a fixed region. Figure 7 demonstrates a nearly direct proportionality relationship between the two. A Linear Regressor line can generalise this dual feature input to provide new estimates of yield for a given rainfall amount. The designed hybrid model predicts crop yield in tonnes per hectare as directed by the rainfall - soil nutrient relationship as depicted in the below graph.

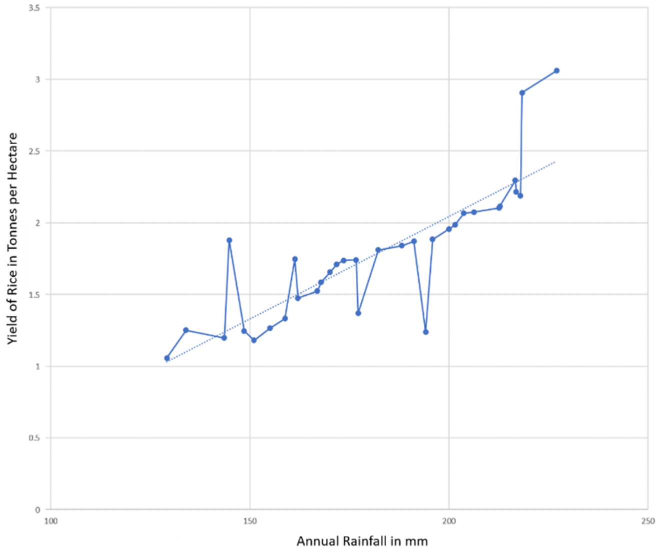

Fig 7. Trend in rainfall intensity versus crop yield volume

V. Evaluation of the Algorithm

Machine learning domains like Neural Networks are highly data centric as every result is obtained by processing input data through several statistically proven strategies. Neural Networks in our scope of research involves such huge input dataframes that are processed through exponential sequencing network of neurons. Unlike conventional statistical modelling, in neural networks, the number of neurons for each layer, the number of epochs and the batch size have been permuted over several scales to obtain the most optimal values for algorithmic calculation.

The proposed RNN model was evaluated on 25% of the dataset tuples as a Test dataset. The predicted values are compared against the recorded label values to measure the error in prediction performance of the algorithm-model. Figure 8 depicts the performance of the model predictions with respect to the recorded data. It can be observed that the predictions curve (the red coloured curve in Figure 8) is able to reasonably trace the observed value curve (the blue coloured curve in Figure 8). Table I explains the trial and error estimation involved in converging to suitable activation functions and optimal number of neurons for the final algorithm-model. The performance of the model with permuted tuples of the hyperparameters can be expressed in terms of the error rate in prediction. Metrics such as RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are employed to gauge the model's error rate.

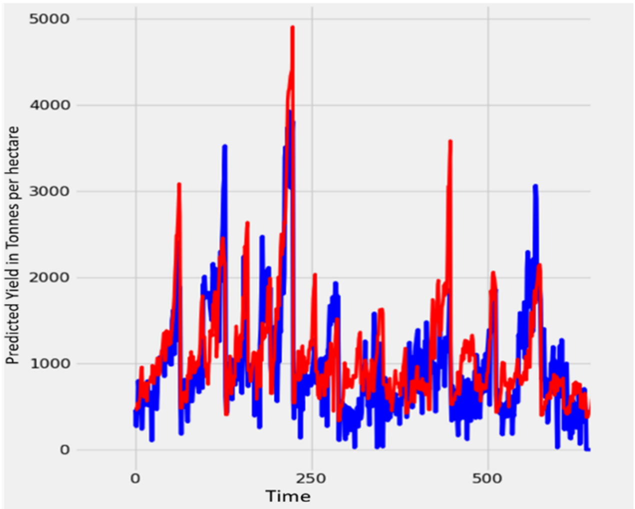

Fig 8. Graph plotted to make a comparison between the actual and predicted yield values

TABLE I
Comparison of Error metrics for various combinations of layers and hyperparameters

INPUT LAYER(no. neurons)	HIDDEN LAYER(no. neurons)	Activation Function(sequence)	BATCH SIZE	EPOCHS	RMSE	MAE
32	24	reLu, reLu, reLu, reLu	4	215	41.497	41.6
32	24	reLu, linear,sigmoid,sigmoid	4	200	39.509	56.8
32	24	linear, tanh, tanh,reLu	10	50	40.829	79.1
32	24	linear, signoid,sigmoid, relu	10	100	42.226	79.3
32	24	reLu, linear,linear,reLu	10	50	43.578	98.1
32	24	sigmoid,sigmoid,sigmoid,sigmoid	10	100	44.643	101.2

VI. Conclusion and Future Work

The algorithm developed introduces a data driven model to predict and forecast crop yield using joint dependencies of soil and climate features. Although there are several techniques existing to obtain rainfall predictions [19], the algorithm discussed in this paper succeeded in emphasising on Rainfall along with the crop yield prediction. This designed model took into account the most relevant environment as well as soil parameters that affect the crop growth, in a way that each of those parameters received equal weightage in the final prediction.

The outcomes of this research can benefit the agriculturists/farmers by knowing the investment capital on the crop to be sown, even before the sowing season begins. The predictive pattern of the algorithm can benefit local self-governments and financial institutions to allocate suitable funds or fiscal loans to farmers.

Although the proposed model produced reasonable error rates, it can be improved further by using larger and newer data inputs that take into account the real-time fluctuations in the climatic and soil conditions due to unfortunate events like cyclones, landslides, floods, etc. Larger dataset could help the model to train and predict with better accuracy. This data centric model reliant on collected data can be sophisticated by introducing Wireless Sensor Networks at real-time to collect and transform data directly to predictive models, unlike the existing process of bulk collection and feed via datasets.

Acknowledgment

We are grateful to M.S. Ramaiah Institute of Technology, Bangalore for providing immense support in guidance and testing of the research project. We are thankful to the work by James McKinsey and Robert Evenson of Yale University for the curation of the Indian Agriculture and Climate dataset. We express our gratitude to Karnataka State Natural Disaster Monitoring Centre for rainfall and weather informatics related to our research.

References

[1] Rasul G, Q. Z. Chaudhry, A. Mahmood and K. W. Hyder, “Effect of. 28-40Temperature Rise on Crop Growth & Productivity”, Pakistan Journal of Meteorology, Volume 8, Issue 15, 2011, pp. 7-8.

[2] Anupama Mahato, “Climate Change and its Impact on Agriculture”, International Journal of Scientific and Research Publications, Volume 4, Issue 4, ISSN 2250-3153, April 2014, pp. 4-5.

[3] Japneet Kaur, “Impact of Climate Change on Agricultural Productivity and Food Security Resulting in Poverty in India”, Università Ca' Foscari Venezia, 2017, pp. 16-18, 23.

[4] Pratap S. Birthal, Md. Tajuddin Khan, Digvijay S. Negi and Shaily Agarwal, “Impact of Climate Change on Yields of Major Food Crops in India: Implications for Food Security”, Agricultural Economics Research Review, Volume 27 (No. 2), pp. 145-155, July-December 2014.

[5] H. Yunis, Y. Bashan, Y. Okon, Y. Henis, “Weather Dependence, Yield Losses, and Control of Bacterial Speck of Tomato Caused by Pseudomonas tomato”, American Phytopathological Society, October 1980, pp. 1-2.

[6] J.P. Powell, S. Reinhard, “Measuring the effects of extreme weather events on yields”, Weather and Climate Extremes 12, pp. 69-79, Elsevier, 2016.

[7] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50., pp. 159-175.

[8] B. Dumont, V. Leemans, Salvador Ferrandis, Bernard bodson, Jean-Perrie Destain, “Assessing the potential of an algorithm based on mean climatic data to predict wheat yield.”, Precision Agriculture, Volume 15, Issue 3, pp. 255-272, June 2014.

[9] Basso B, Bodson B, V. Leemans, B. Bodson, J-P Destain, M-F Destain, “A comparison of within season yield predictions algorithm based on crop model behaviour analysis”, Agricultural and Forest Meteorology, Volume 204, pp. 10-21, May 2015.

[10] Stanley A Changnon. “Prediction of corn and soya bean yields using weather data”, CHIAA Research Report No. 22, Crop-Hail Insurance Actuarial Association, February 1965, pp. 6-10.

[11] Betty. J, Shem G Juma, Everline. O, “On the Use of Regression Models to Predict Tea Crop Yield Responses to Climate Change: A Case of Nandi East, Sub-County of Nandi County, Kenya”, Assesing the Value of Systematic Cycling in a Polluted Urban Environment, Climate, Volume 5, Issue 3, July 2017, pp. 5.

[12] Christian Baron and Mathieu Vrac, Oettli. P, Sultan. B, “Are regional climate models relevant for crop yield prediction in West Africa?”, Environmental Research Letters, Volume 6, 2011, pp. 2-6.

[13] https://www.ksndmc.org/ReportHomePage.aspx

[14] http://drdpat.bih.nic.in/PA-Table-10-Karnataka.htm

[15] https://facebook.github.io/prophet/docs/quick_start.html/ ; last visited: 21 April, 2018

[16] https://blog.exploratory.io/is-prophet-better-than-arima-for-forecasting-time-series-fa9ae08a5851

[17] https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials/ ; last visited: 10 April, 2018

[18] https://www.statisticshowto.com/exponential-smoothing/ ; last visited: 21 April, 2018

[19] Mr. Dhawal Hirani, Dr. Nitin Mishra, “A Survey on Rainfall Prediction Techniques”, International Journal of Computer Application (2250-1797), Volume 6- No.2, March-April 2016, pp. 28-40.